KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0341

PROJECTION PURSUIT BASED MEASURES OF ASSOCIATION

by
C. CROUX
P. FILMOZER

D/2003/2376/41

Projection Pursuit based Measures of Association

C. Croux [*] P. Filzmoser [†]

K.U. Leuven Vienna University of Technology

Abstract: In this paper we present measures of association between two multivariate stochastic variables X and Y based on the idea of projection pursuit. The association measure is defined as the maximal value that a bivariate correlation index between one-dimensional projections of X and Y can attain. Taking the Pearson correlation as projection index results in the first canonical correlation coefficient. Other projection indices, like Spearman's rank correlation, will yield more non-parametric or robust measures of association. Robustness of the association measures is studied by computing influence functions. By means of a simulation study, the stability and precision of the different measures has been compared. Using the association measures, a permutation test for independence of two multivariate vectors is formulated. Such tests can be more powerful for detecting non-linear relationships.

Keywords: Association measures, Correlation measures, Influence function, Multivariate test for independence, Projection pursuit, Rank correlation, Robustness.

[*]Department of Applied Economics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium. Email: christophe.croux@econ.kuleuven.ac.be

[†]Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria. Email: P.Filzmoser@tuwien.ac.at

1 Introduction

Correlation between two univariate variables U and V can be measured in several ways. The correlation coefficients of Pearson, Spearman and Kendall, among others, are standard tools in statistical practice. For measuring the degree of association between two multivariate vectors X and Y much less variety is existing. We introduce a class of measures of association between multivariate vectors based on the idea of *projection pursuit*. Projection pursuit aims at finding "interesting" projections (typically one- or two-dimensional) of a multivariate data set, where interestingness is measured by a projection index (see e.g. Huber 1985).

Suppose that X is a p-dimensional random variable and Y a q-dimensional random variable, with $p \leq q$. A measure of multivariate association between X and Y can be defined by looking for linear combinations $a^t X$ and $b^t Y$ of the original variables having maximal association. Expressed in mathematical terms, we seek for a measure

$$\rho_R(X,Y) = \max_{\alpha,\beta} R(\alpha^t X, \beta^t Y), \tag{1.1}$$

where R is a measure of association between univariate variables. Using the the projection pursuit terminology, R is the *projection index* to maximize. Depending upon the bivariate measure R used in the above definition, different measures of association between X and Y are obtained. Taking for R the classical Pearson correlation measure results in the first *canonical correlation coefficient* (see e.g. Johnson and Wichern, 1998). Other choices of R yield measures ρ_R having different properties. Bivariate correlation measures which will be considered are Spearman's rank correlation, Kendall's tau, the correlation median (Falk 1998), a bivariate M-estimator (Huber 1981), and the bivariate Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1985).

The vectors $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ yielding the maximum in (1.1) will be called the weighting vectors. To identify them uniquely (upto a sign) we impose a unit norm restriction. So,

$$(\alpha_R(X,Y), \beta_R(X,Y)) = \operatorname*{argmax}_{\|\alpha\|=1, \|\beta\|=1} R(\alpha^t X, \beta^t Y). \tag{1.2}$$

We will call α_R and β_R the *weighting vectors.* They indicate the contribution of every single component of X and Y in the construction of the indices $\alpha_R^t X$ and $\beta_R^t Y$ giving maximal association.

In Section 2 of the paper we formally define the association measures and give some basic properties. Section 3 shows us influence functions of the association measures and of the weighting vectors. It will be shown that using a projection index R having a bounded influence function yields bounded influence for ρ_R, while this does not longer hold for the weighting vectors α_R and β_R. An approximative algorithm for computing the measure is presented in Section 4. In Section 5 a permutation test for testing independence between two multivariate stochastic variables using the projection pursuit based association measure is introduced. Several examples will be given here. Section 6 presents a simulation study to investigate the robustness and efficiency of different estimators of association, and to study the power of the permutation test for detecting linear and monotone non-linear relations between X and Y.

2 Definitions and Basic Properties

Denote R the projection index we will maximize. The correlation measure R needs so verify the following properties, where (U, V) stands for any pair of univariate variables:

(i) $R(U, V) = R(V, U)$

(ii) $R(aU + b, cV + d) = \operatorname{sign}(ac)R(U, V)$ for all $a, b, c, d \in I\!R$

(iii) U and V independent $\Rightarrow R(U, V) = 0$

Note that condition (ii) gives $R(-U, V) = -R(U, V)$. Therefore $\rho_R \geq 0$ and ρ_R can also be defined as

$$\rho_R(X, Y) = \max_{\alpha,\beta} |R(\alpha^t X, \beta^t Y)|.$$

Condition (iii) allows to test for independence using the association measure ρ_R. Indeed, if X and Y are independent, then every $\alpha^t X$ is independent of every other $\beta^t Y$, yielding $\rho_R(X, Y) = 0$. The equivariance property (ii) ensures the association measures to be invariant under affine transformations. Indeed, for any non-singular matrices A and B and vectors a and b one has

$$\rho_R(AX + b, BY + b) = \rho_R(X, Y).$$

The weighting vectors are affine equivariant in the sense that

$$\alpha_R(AX + b, BY + b) = (A^t)^{-1}\alpha_R(X, Y)/\|(A^t)^{-1}\alpha_R(X, Y)\|$$

and similarly for β_R. The normalization in the above formula is implied by the unit norm restriction on the weighting vectors. This restriction is in a sense arbitrary, and could be replaced by imposing a unit variance for $\alpha_R^t X$ and $\beta_R^t Y$. The latter would however require the choice of a suitable measure of variance, and yields in any case besides a scaling factor, the same weighting vectors. Now we will briefly review the definitions of several bivariate association measures R.

Pearson's correlation: This classical measure for linear association is defined as

$$\operatorname{Corr}(U,V) = \frac{\operatorname{Cov}(U,V)}{\sqrt{\operatorname{Var}(U)\operatorname{Var}(V)}}. \tag{2.1}$$

The maximization problem in (1.1) can now be solved explicitely, since it corresponds to the definition of the first canonical correlation coefficient. We have that ρ^2_{Corr} is given by the largest eigenvalue of the matrix

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}, \tag{2.2}$$

where $\Sigma_{xx} = \operatorname{Cov}(X), \Sigma_{yy} = \operatorname{Cov}(Y),$ and $\Sigma_{xy} = \operatorname{Cov}(X,Y)$ The weighting vector α_{Corr} is then the associated unit norm eigenvector of the matrix (2.2) while β_{Corr} is proportional to $\Sigma_{yy}^{-1}\Sigma_{yx}\alpha_{\text{Corr}}$. Existence of ρ_{Corr} requires existence of second moments, while the other measures to be discussed do not require any existence of moments.

Spearman and Kendall correlation: These famous measures are based on ranks and signs, and are often called non-parametric correlation measures. The Spearman rank correlation is defined as

$$R_S(U,V) = \operatorname{Corr}(\operatorname{rank}(U), \operatorname{rank}(V)),$$

where $\operatorname{rank}(u) = F_U(u)$, with F_U the cumulative distribution function of U, stands for the population rank of u. The definition of Kendall's tau is

$$R_K(U,V) = E[\operatorname{sign}((U_1 - U_2)(V_1 - V_2))]$$

where (U_1, V_1) and (U_2, V_2) are two independent copies of (U,V). Estimators of the population correlation measures are simply given by the sample counterparts. For example, from an i.i.d. sample $(u_1, v_1), \ldots, (u_n, v_n)$ we can compute the sample version of $R_K(U,V)$:

$$\hat{R}_K = \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{sign}((u_i - u_j)(v_i - v_j)).$$

Correlation derived from bivariate scatter matrices: A scatter matrix C can be seen as a robust alternative to the classical covariance matrix. An important example are M-estimators of Maronna (1976). Given a 2-dimensional variable $Z = (U,V)^t$, the M-location $\mu(Z)$ and M-scatter matrix $C(Z)$ are implicitly defined as solutions of the equations

$$\begin{aligned} \mu &= E\Big[w_1\left((Z-\mu)^t C^{-1}(Z-\mu)\right) Z\Big]/E\Big[w_1\left((Z-\mu)^t C^{-1}(Z-\mu)\right)\Big] \\ C &= E\Big[w_2\left((Z-\mu)^t C^{-1}(Z-\mu)\right)(Z-\mu)(Z-\mu)^t\Big] \end{aligned}$$

where μ is a bivariate vector and C is a symmetric positive definite two-by-two matrix. Furthermore w_1 and w_2 are specified weight functions. We focus on Huber's M-estimator, obtained by taking $w_1(d^2) = \max(1, \tau/d)$ and $w_2(d^2) = c \max(1, (\tau/d)^2)$ with $\tau = \chi^2_{2,0.9}$ the 10% upper quantile of a chi-squared distribution with 2 degrees of freedom and c selected to obtain a consistent estimator of the covariance matrix at normal distributions (Huber 1981).

M-estimators of scatter can be considered as (iteratively) reweighted covariance matrices, and are easy to compute. However, they loose a lot of robustness under huge amounts of contamination. Therefore high breakdown multivariate scatter matrices are often used as a more resistant measure of multivariate scatter. A popular estimator is the minimum covariance determinant estimator (Rousseeuw 1985, and Rousseeuw and Van Driessen 1999, for a fast algorithm). The minimum covariance determinant (MCD) estimator is determined by that subset of observations of size h which minimizes the determinant of the sample covariance matrix, computed from only these h points. The location estimator is the average of these h points, whereas the scatter estimator is proportional to their covariance matrix. As a compromise between robustness and efficiency, selected $h = \lfloor 0.75n \rfloor$ will be taken. The population version of the MCD scatter matrix is formally defined in Butler, Davies, and Jhun (1993).

For a review of estimators of multivariate location and scatter we refer to Maronna and Yohai (1998). The correlation measure associated with a bivariate scatter matrix $C(Z) \equiv C(U, V)$ is then simply given by

$$R_C(U, V) = \frac{C_{12}(U, V)}{\sqrt{C_{11}(U, V) C_{22}(U, V)}}.$$

The correlation measures based on Huber's M and the MCD bivariate scatter matrices will be denoted by R_M and R_{MCD}.

The correlation median:

Besides the so-called non-parametric correlations and correlation derived from scatter matrices, several other robust measures of correlation have been proposed in the literature (see Shevlyakov and Vilchevski 2002, Chapter 7, for an overview). One of the most simple one is the correlation median (Falk 1998), obtained by replacing averages by medians in the definition of Pearson's correlation:

$$R_{\mathrm{Comed}}(U, V) = \frac{\mathrm{med}(U - \mathrm{med}U)(V - \mathrm{med}V)}{\mathrm{med}|U - \mathrm{med}U| \; \mathrm{med}|V - \mathrm{med}V|}.$$

Note that all considered measures of correlation R verify condition (i), (ii), and (iii). Moreover, they all return values in the interval [-1,1] with the exception of the correlation median which is not always guaranteed to be smaller than 1.

It is important to realize that different measures of correlation R represent different population quantities. Consider the bivariate normal distribution

$$\Phi_\rho = N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{2.3}$$

and let R be a measure of bivariate correlation. Define then the function $\kappa_R : [-1,1] \rightarrow I\!R$ by

$$\kappa_R(\rho) = R(\Phi_\rho) \quad \text{for any } -1 < \rho < 1. \tag{2.4}$$

For example, for Spearman and Kendall correlation it is known that

$$\kappa_S(\rho) = \frac{6}{\pi}\sin^{-1}(\rho/2) \quad \text{and} \quad \kappa_K(\rho) = \frac{2}{\pi}\sin^{-1}(\rho).$$

Since the MCD and M-estimators are consistently estimating the shape of normal distributions, they estimate the same quantity as Pearson's correlation: $\kappa_{\text{MCD}}(\rho) = \kappa_M(\rho) = \rho$. For the correlation median, it is shown in Falk (1998) that $\kappa_{\text{Comed}}(\rho)$ is the unique solution of the following equation in κ:

$$\int_0^{\frac{1}{\pi}\cos^{-1}(-\rho)} \exp\left(\frac{-2\left\{\Phi^{-1}\left(\frac{3}{4}\right)\right\}^2 \kappa}{\cos(\pi u) + \rho}\right) du = \frac{1}{2}$$

and can be computed numerically for every $-1 < \rho < 1$.

In the next Section the influence function of the association measures and the weighting vectors will be computed at the multivariate normal model. The assumption

(iv) κ_R is strictly increasing and differentiable,

will be imposed. It is immediate to see that (iv) holds for all considered estimators (for the correlation median, a proof is given in Falk 1998).

3 Influence Functions

In this section we will compute the influence functions (IFs) of the previously defined association measure and weighting vectors. The IF gives the influence that an observation x has on a functional Q at a distribution H. If we denote a point mass distribution at x by Δ_x and write $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_x$ then the IF is given by

$$IF(x, Q, H) = \frac{\partial}{\partial\varepsilon} Q(H_\varepsilon)\Big|_{\varepsilon=0}.$$

(see Hampel at al. 1986). By convention, $Q(H) \equiv Q(X,Y)$ when $(X,Y) \sim H$, for any statistical functional Q. Take $(X,Y) \sim H$ a $(p+q)$-dimensional distribution function. Recall that the statistical functionals of interest here are defined as

$$(\alpha_R(H), \beta_R(H)) = \underset{\|\alpha\|=1, \|\beta\|=1}{\operatorname{argmax}} R(\alpha^t X, \beta^t Y) \tag{3.1}$$

and

$$\rho_R(H) = R(\alpha_R(H)^t X, \beta_R(H)^t Y). \tag{3.2}$$

The influence function will be computed at the normal model distribution

$$H_0 = N(0, \Sigma) = N\left(0, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right),$$

where the location is without lost of generality taken to be zero, due to the translation invariance of the functionals. We will also suppose throughout the paper that Σ has full rank. At this model distribution we have that

$$R(\alpha^t X, \beta^t Y) = \kappa_R\left(r(\alpha, \beta)\right) \tag{3.3}$$

where

$$r(\alpha, \beta) = \frac{\alpha^t \Sigma_{xy} \beta}{\sqrt{\alpha^t \Sigma_{xx} \alpha}\sqrt{\beta^t \Sigma_{yy} \beta}} = \operatorname{Corr}(\alpha^t X, \beta^t Y).$$

Since κ_R is supposed to be strictly increasing, it follows that the functionals $\alpha_R(H_0)$ and $\beta_R(H_0)$ defined in (3.1) are the same for all correlation measures verifying (iv). Taking $R = \operatorname{Corr}$ yields then immediately that

$$\alpha_R(H_0) := \alpha_1/\|\alpha_1\| \text{ and } \beta_R(H_0) := \beta_1/\|\beta_1\| \tag{3.4}$$

where α_1 and β_1 are the first canonical correlations at H_0. The calculus for canonical correlation analysis can be found in textbooks like (Johnson and Wichern 1998, Chapter 10, or Rencher 1998, Chapter 8). Solving the maximization problem in (3.1) using Lagrange multiplicators reveals that both (α_1, β_1) and $(\alpha_R(H_0), \beta_R(H_0))$ satisfy the following system

$$\left\{ \begin{array}{lcl} \Sigma_{xy}\beta & = & r(\alpha,\beta)\Sigma_{xx}\alpha \\ \Sigma_{yx}\alpha & = & r(\alpha,\beta)\Sigma_{yy}\beta. \end{array} \right. \tag{3.5}$$

Furthermore, it follows from (3.2) that

$$\rho_R(H_0) = R(\alpha_R(H_0)^t X, \beta_R(H_0)^t Y) = \kappa_R\left(r(\alpha_1, \beta_1)\right) = \kappa_R(\rho_1), \tag{3.6}$$

where ρ_1 stands for the first population canonical correlation.

A fairly simple expression for the influence functions can now be derived. The IF for the weighting vectors will be expressed in terms of higher order population canonical correlations. Recall from standard canonical correlation analysis (at the population level) that the squared canonical correlations $\rho_1^2, \ldots, \rho_p^2$ are the eigenvalues of (2.2) in descending order. Then $\alpha_1, \ldots, \alpha_p$ are the associated eigenvectors with $\alpha_k^t \Sigma_{xx} \alpha_k = 1$ and $\alpha_k^t \Sigma_{xx} \alpha_j = 0$, for every $1 \leq k < j \leq p$. The variables $U_k = \alpha_k^t X$ are called the canonical variates and are mutually uncorrelated. Similarly, $\beta_1, \ldots, \beta_q$ are eigenvectors of

$$\Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy},$$

such that $\beta_k^t \Sigma_{yy} \beta_k = 1$, for $k = 1, \ldots, q$, and $\beta_k^t \Sigma_{yy} \beta_j = 0$ for $k \neq j$.

Theorem 1 *Let R be a correlation measure satisfying conditions (i) upto (iv), and take $H_0 = N(0, \Sigma)$ a multivariate normal distribution. Let $\rho_1 > \ldots > \rho_p > 0$ be the population canonical correlations and $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$ be the population canonical vectors. Let for every $j = 1, \ldots, p$ the canonical variates be $u_j = x^t \alpha_j$ and $v_j = y^t \beta_j$. Then we have that the influence function of the association measure is given by*

$$IF((x, y), \rho_R, H_0) = IF((u_1, v_1), R, \Phi_\rho), \tag{3.7}$$

the IF for the weighting vectors by

$$\begin{aligned} IF\left((x, y), \alpha_R, H_0\right) &= \sum_{j=2}^{p} \frac{1}{\rho_1^2 - \rho_k^2} \left\{ IF_1((u_1, v_1), R, \Phi_\rho)\, \rho_1 u_k \right. \\ &\left. + IF_2((u_1, v_1), R, \Phi_\rho)\, \rho_k v_k \right\} \left(I - \frac{\alpha_1 \alpha_1^t}{\|\alpha_1\| \|\alpha_1\|} \right) \frac{\alpha_k}{\|\alpha_1\| \kappa_R'(\rho_1)} \end{aligned} \tag{3.8}$$

and

$$\begin{aligned} IF\left((x, y), \beta_R, H_0\right) &= \sum_{j=2}^{q} \frac{1}{\rho_1^2 - \rho_k^2} \left\{ IF_1((u_1, v_1), R, \Phi_\rho)\, \rho_1 v_k \right. \\ &\left. + IF_2((u_1, v_1), R, \Phi_\rho)\, \rho_k u_k \right\} \left(I - \frac{\beta_1 \beta_1^t}{\|\beta_1\| \|\beta_1\|} \right) \frac{\beta_k}{\|\beta_1\| \kappa_R'(\rho_1)} \end{aligned} \tag{3.9}$$

where Φ_ρ is a bivariate normal distribution with correlation ρ, and $\rho_k := 0$ for $k > p$. The partial derivates w.r.t. the first, respectively the second component, of $IF((u_1, v_1), R, \Phi_\rho)$ have been denoted by IF_1 and IF_2.

Proof of Theorem 1: see Appendix.

Theorem 1 shows that the influence function of the projection index R determines the shape of the IF for the multivariate association measure ρ_R and the weighting vectors. While a bounded $IF((u_1, v_1, R, \Phi_\rho)$ ensures a bounded influence function for ρ_R, this is no longer true for the weighting vectors. Indeed, if u_k or v_k tend to infinity (for a $k \geq 2$), the influence function for the weighting vector goes beyond all bounds. This happens even when the bivariate correlation measure has an influence function redescending to zero for large values of (u_1, v_1). Note that an unbounded influence function means that if there is a small amount ε of badly placed contamination, the change in the value of the functional will be disproportionally large with respect to the level of contamination. It does not mean that the functional breaks down or explodes in presence of small amounts of outliers.

Furthermore, since the partial derivatives of $IF((u_1, v_1, R, \Phi_\rho)$ appear in the IF for the weighting vectors, it is necessary to take a projection index R having a smooth influence function. In particular, discontinuities in $IF((u_1, v_1, R, \Phi_\rho)$ yield highly unstable estimates of the weighing vectors. In Figures 1a upto 1f, graphs of $IF((u, v, R, \Phi_\rho)$ are pictured for $\rho = 0.5$ and for all projection indices considered in this paper.

For the Pearson correlation, Devlin et al (1975) showed that

$$IF((u,v), R_{\text{Corr}}, \Phi_\rho) = uv - \rho \frac{u^2 + v^2}{2}.$$

It is then easy to verify that the formulas of Theorem 1 correspond with known expressions for the influence function of the canonical correlations and canonical vectors derived from the sample covariance matrix. The first one to derive the IF for classical canonical correlations was, upto our knowledge, Romanazzi (1992). His derivation for the IF relies on the eigenvalue analysis of the matrix (2.2), and does not use projection-pursuit definitions (1.1) and (1.2) taking Pearson correlation as projection index. Figure 1a shows that using Pearson's correlation as projection index yields a highly non robust procedure.

The IF functions for Spearman correlation and Kendall's τ are pictured in Figures 1b and 1c. Although computation of these influence functions is quite straightforward, we could not retrieve expressions for these IFs from the literature (except for $\rho = 0$). From the Figures we see that the IFs are very smooth and bounded. The form of these IFs suggest good robustness behavior against small model departures for these correlation measures.

Formulas for the IFs for the correlation measures derived from any affine equivariant bivariate scatter are immediately obtained combining Lemma 2 in Croux and Haesbroeck

(2000) and expressions for the influence function of an off-diagonal element of the scatter matrix. The latter are well-known for M-estimators (e.g. Hampel et al 1986, Chapter 5), and have been obtained by Croux and Haesbroeck (1999) for the MCD. Figure 1d shows the IF for R_M: in the center of the distribution it has the same form as the IF of the classical correlation, but in the extremes it is bounded above. In a way, its form is similar to the IF of R_S and R_K. For R_{MCD}, it is seen from Figure 1e that the IF is not only bounded, but also redescending abruptly to zero for far away outliers. The jumps in the IF of R_{MCD} indicate that the weighting vectors will be unstable in presence of small amounts of contamination, and that the asymptotic behavior of the estimates of the weighting vector will be non-standard (as for the Least Median of Squares regression estimator).

Finally, standard influence function calculation gives for the IF of the correlation median

$$IF((u,v), R_{\text{Comed}}, \Phi_\rho) =$$

$$\frac{1}{\Phi^{-1}(\frac{3}{4})^2}\left(\frac{\text{sign}(uv-\rho_M)}{2k(\rho_M)} - \frac{\kappa_{\text{Comed}}(\rho)}{\Phi^{-1}(\frac{3}{4})}\left(IF(u,\text{MAD},\Phi)+IF(v,\text{MAD},\Phi)\right)\right)$$

where k is the density of XY, and $\rho_M = \kappa_{\text{Comed}}(\rho)\Phi^{-1}(\frac{3}{4})^2$. From Figure 1f one can see that although the IF is bounded, it has many jumps indicating poor local robustness properties.

Influence functions measure the local robustness of an estimator, i.e. robustness w.r.t. small amounts of contamination. To measure the robustness of a correlation measure under larger amounts of outliers, the maxbias curve is more appropriate. Research on maxbias curves for correlation measures is currently ongoing.

4 Computational Aspects

The association measures have been defined for arbitrary distributions H in (1.1). The sample counterpart of $\rho_R(H)$ is then given by $\hat{\rho}_R := \rho_R(H_n)$, where H_n is the empirical distribution function of a sample $(x_1, y_1), \ldots, (x_n, y_n)$ from H. The computation of

$$\hat{\rho}_R = \max_{\|\alpha\|=1, \|\beta\|=1} R_n\left((\alpha^t x_1, \beta^t y_1), \ldots (\alpha^t x_n, \beta^t y_n)\right), \tag{4.1}$$

where R_n is the sample version of R, is not straightforward. An objective function needs to be maximized over a $(p+q)$ dimensional space under constraints. Moreover, these objective functions do not need to be smooth. In the special case of $p = 1, q = 2$ it is possible to visualize the shape of this objective function, since without lost of generality the search space can be parameterized as $\{(1, (\cos\theta, \sin\theta)) | -\pi < \theta \leq \pi\}$. In Figure 2 it is shown how the value of the objective function varies with θ. It is immediately seen that the objective

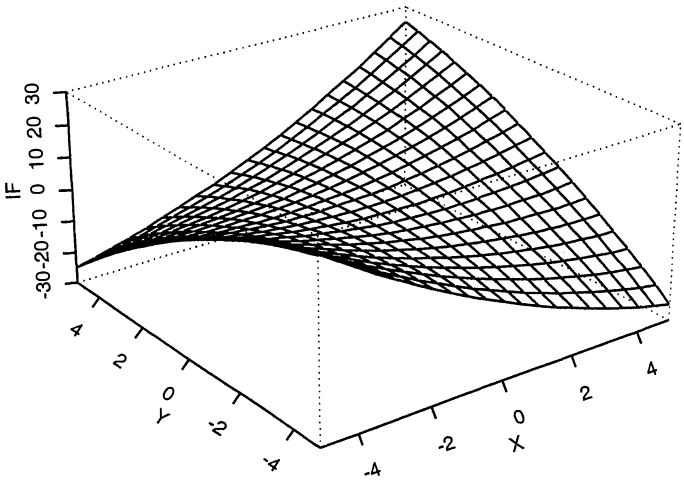


Figure 1a: IF of Pearson's correlation at $\Phi_{0.5}$.

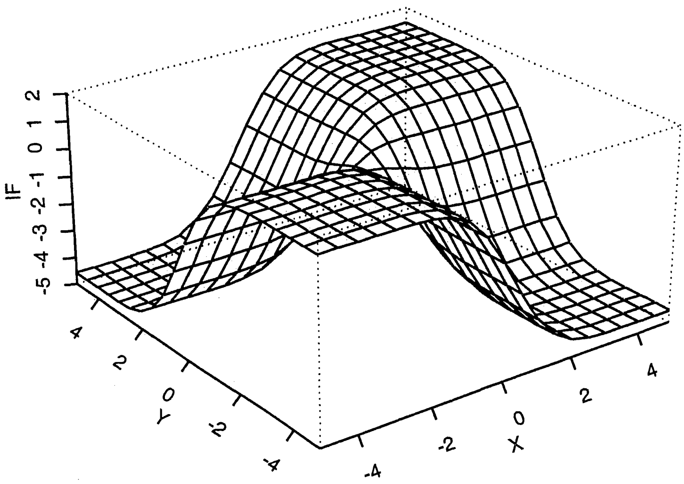


Figure 1b: IF of Spearman's rank correlation at $\Phi_{0.5}$.

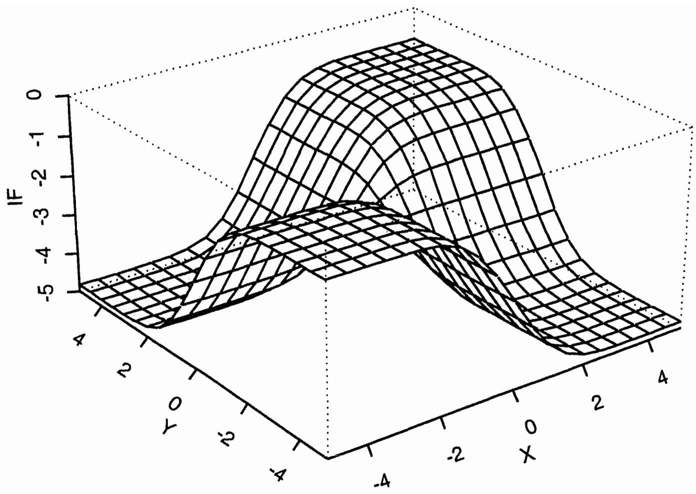


Figure 1c: IF of Kendall's τ *at* $\Phi_{0.5}$.

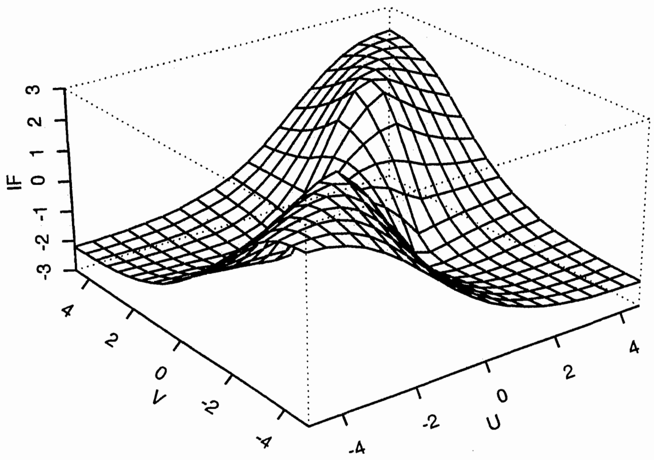


Figure 1d: IF of correlation based on a bivariate Huber's M scatter matrix at $\Phi_{0.5}$.

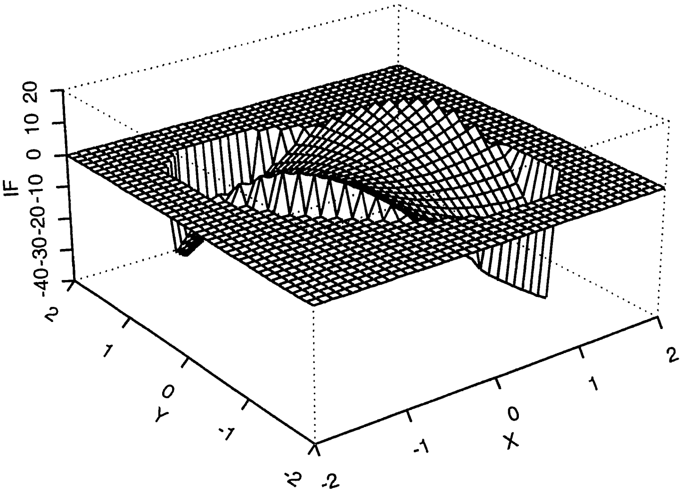


Figure 1e: IF of correlation based on a bivariate MCD scatter matrix at $\Phi_{0.5}$.

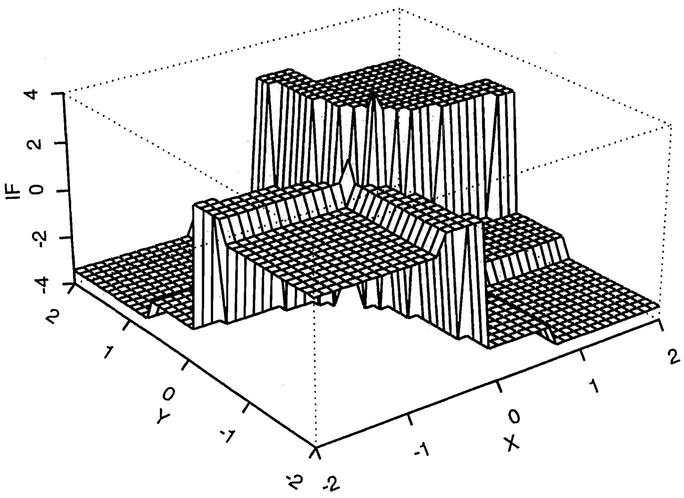


Figure 1f: IF of the correlation median at $\Phi_{0.5}$.

functions for $\hat{\rho}_{\text{Comed}}$ and to a lesser extend for $\hat{\rho}_{\text{MCD}}$ behave erratically, already indicating instability of these estimators. The objective functions for $\hat{\rho}_{\text{Corr}}$ and $\hat{\rho}_M$ are very smooth, and would allow for local gradient search. For the non-parametric correlation measures $\hat{\rho}_S$ and $\hat{\rho}_K$, the objective function is continuous, but has a non-continuous derivative.

When performing projection-pursuit principal components analysis (Li and Chen 1985) similar computational problems arise, and approximative algorithms were developed. The basic algorithm we will use is inspired on Croux and Ruiz-Gazen (1996). Denote $R_n(\alpha, \beta)$ as shorthand notation for the objective function in (4.1) to maximize. The (absolute value of the) objective function will only be evaluated at (α, β) belonging to the finite set

$$D_n = \{(a_i, b_j)|a_i = \frac{x_i - m_X}{\|x_i - m_X\|} \text{ and } b_j = \frac{y_j - m_Y}{\|y_j - m_Y\|} \text{ for } i, j = 1, \ldots, n\}$$

with m_X and m_Y being a location estimator of the x and the y-observations, e.g. the coordinatewise median or the spatial median. The set D_n contains possible interesting directions, in the sense that the (α, β) we are looking at point in the direction of the data. Since D_n has n^2 elements an exhaustive search over it is only feasible for very small values of n. Therefore the following greedy search procedure over D_n will be followed:

- Select randomly two indices (i_0, j_0) and compute

  $$\max_j |R_n(a_{i_0}, b_j)| \text{ and } \max_i |R_n(a_i, b_{j_0})|.$$

  Denote $j(i_0)$ and $i(j_0)$ the indices corresponding to the above maxima.

- If $|R_n(a_{i_0}, b_{j(i_0)})| > |R_n(a_{i(j_0)}, b_{j_0})|$, then approximate $\hat{\rho}_R$ by $\max_i |R_n(a_i, b_{j(i_0)})|$. Otherwise, approximate $\hat{\rho}_R$ by $\max_j |R_n(a_{i(j_0)}, b_j)|$.

- The estimates of the weighting vectors are then given by the (a_i, b_j) corresponding to the above maxima.

We do not claim that the proposed fast algorithm is optimal, but it is simple and fast to carry out. It can be directly applied for any given projection measure R, without additional calculations to make. The quality of the approximation of the algorithm is mainly determined by the sample size n. For samples from continuous distributions the algorithm tends to the exact $\hat{\rho}_R$ for n tending to infinity. For smaller values of n, it is advantageous to make additional evaluation of the objective functions at randomly generated values of α and β (similar as in Boente, Pires, and Rodrigues 2002). As for projection-pursuit principal components analysis (PCA), there is still a challenge for computational statisticians to provide

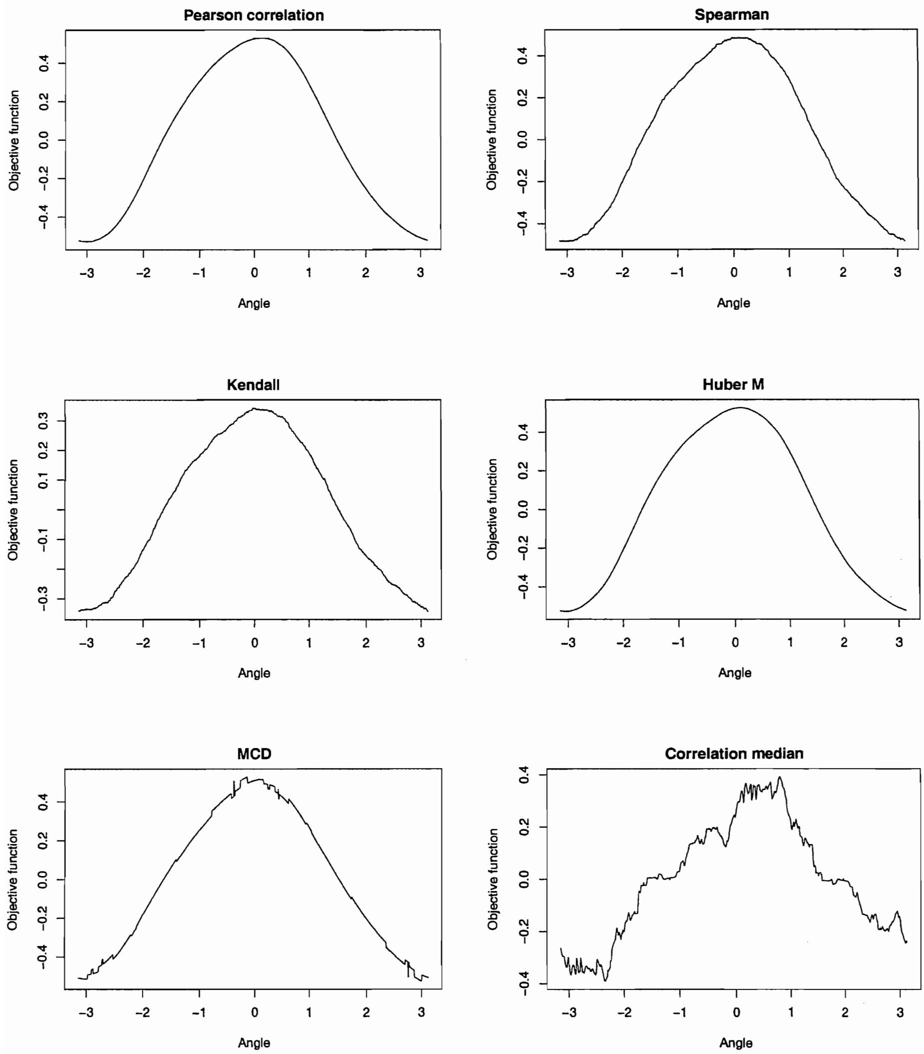


Figure 2: Shape of the objective functions of the association measures for the projection indices Pearson correlation (Corr), Spearman's rank correlation (S), Kendall's τ (K), correlation based on Huber's bivariate M-estimator (M) and on the bivariate MCD, and the correlation median (Comed) in case $p = 1, q = 2$.

Table 1: MAPE of the fast algorithm for computing the association measures at simulated data sets with size $n = 100$, and $p = 1, q = 2$. Several projection indices were considered.

R	Corr	S	K	M	MCD	Comed
Normal data	0.0002	0.0027	0.0035	0.0003	0.0060	0.0221
10% intermediate outliers	0.0003	0.0028	0.0035	0.0005	0.0061	0.0242
10% extreme outliers	0.0145	0.0024	0.0033	0.0076	0.0420	0.0271

fast, robust and accurate algorithms to compute (4.1). The following numerical experiment indicates that the basic algorithm we proposed ahead yields quite satisfactory results.

In case $p = 1, q = 2$ it is possible to compute $\hat{\rho}_R$ upto any desired accuracy by a simple grid search over angles in the interval $[-\pi, \pi]$. We simulated $m = 1, \ldots, M = 1000$ samples of size $n = 100$ and computed for each generated sample $\hat{\rho}_R^m$ and the approximation $\tilde{\rho}_R^m$ using the algorithm outlined above. Then the Mean Absolute Percentwise Error

$$\text{MAPE} = \frac{1}{M} \sum_{m=1}^{M} \left| \frac{\hat{\rho}_R^m - \tilde{\rho}_R^m}{\hat{\rho}_R^m} \right|$$

has been computed. This experiment was repeated for samples containing 10% of intermediate and extreme outliers. In presence of outliers the objective function is expected to have more local maxima. For an explicit description of the sampling scheme, we refer to Section 6, were the same design is used for a simulation study. The results are presented in Table 1. We see that the MAPE are in most cases smaller than 1%, indicating the good performance of the algorithm in this setting. An exception is the association measure based on the correlation median, where the MAPE is larger. This is due to the many jumps we observed in the objective function for Comed, see Figure 2. Also, with 10% of extreme outliers in the data the MAPE of Corr and MCD become worse.

Another situation were $\hat{\rho}_R$ can be computed exactly is when R is taken to be the usual correlation coefficient. Then $\hat{\rho}_R$ is computed from the data as the first eigenvalue of (2.2). We generated 1000 samples of size $n = 100$ from a normal distribution with $p = 3$, $q = 3$ and Σ the identity matrix except for $\Sigma_{14} = \Sigma_{41} = 0.5$, $\Sigma_{25} = \Sigma_{52} = 0.3$ and $\Sigma_{36} = \Sigma_{63} = 0.1$. The MAPE was computed as before resulting in MAPE=0.0151. The MAPE remains small, but is expected to increase in higher dimensions.

Using the fast algorithm, differences in computation time finally depend on the selected bivariate association measure R. There are huge differences in computing time: Pearson, Spearman and the correlation median are very fast to compute and ask only $O(n)$

or $O(n \log n)$ computation time. The bivariate M-estimator is computed according to an iterative scheme, starting from the covariance matrix, and takes already some more time. For the bivariate MCD estimator the Fast-MCD program of Rousseeuw and Van Driessen (1999) implemented in *S-Plus* (1999) was used (here an additional reweighting step is performed, improving the stability of the method considerably): while one evaluation of the MCD is indeed fast to carry out, multiple evaluation of the projection index makes computation of $\hat{\rho}_{\mathrm{MCD}}$ time consuming in comparison with the other measures. Finally, note that the $O(n^2)$ computation time of R_K makes Kendall's τ much less attractive than Spearman's rank correlation in terms of computational complexity.

5 Examples and Permutation Tests for Independence

Description of the "Diabetes data": The "Diabetes data" (Andrews and Herzberg 1985, page 215) measure for a group of $n = 76$ persons the variables Relative Weight, Fasting Plasma Glucose, Glucose Intolerance, Insulin Response to Oral Glucose, and Insulin Resistance. It is of medical interest to establish a relation between the first two (X_1, X_2) and the last three variables (Y_1, Y_2, Y_3).

We standardized the data (centering by the median, scaling by the median absolute deviation) in order to make the weighting vectors comparable for different variables. The association measures, as well as the weighting vectors have been computed for several choices of the projection index R and are given in Table 2. The measures of association indicate weak, but probably significant association between the two sets of variables. Note that the ρ_R measure different characteristics of the data, hence their numerical values are not directly comparable. Looking at the weighting vectors reveals that X_1 and Y_1 contribute most to the index $\hat{\rho}_R$.

Description of the "HIV data": For 36 HIV-positive newborn children their CD45RA T cell counts and CD45RO T cell counts were measured at birth (X_1 and X_2) and after 24 weeks of treatment (Y_1 and Y_2) by a ritonavir therapy. These data have been used by Randles (2000). As in the previous example, we first standardized the data. The association measures, as well as the weighting vectors have been computed for several choices of the projection index R and are given in Table 3. The association measures indicate rather strong association between the two sets of variables. Note that the correlation median can indeed take values larger than one. The weighting vectors indicate now that variables X_2 and Y_2 contribute most to this association. Note that the values of the weighting coefficients is quite variable

Table 2: Association measures and weighting vectors for the Diabetes data for several choices of the projection index R

R	Corr	S	K	M	MCD	Comed
$\hat{\rho}_R$	0.49	0.51	0.38	0.53	0.77	0.53
$\hat{\alpha}_1$	-0.26	-0.26	-0.32	-0.26	-0.37	-0.37
$\hat{\alpha}_2$	0.97	0.97	0.95	0.97	0.93	0.93
$\hat{\beta}_1$	0.97	0.99	0.99	0.97	0.89	0.87
$\hat{\beta}_2$	-0.14	0.10	0.10	-0.14	0.42	0.13
$\hat{\beta}_3$	0.18	-0.07	-0.07	0.18	0.16	-0.48

Table 3: Association measures and weighting vectors for the "HIV data" for several choices of the projection index R

R	Corr	S	K	M	MCD	Comed
$\hat{\rho}_R$	0.86	0.84	0.65	0.81	0.86	1.30
$\hat{\alpha}_1$	-0.16	-0.16	-0.16	-0.16	-0.72	-0.16
$\hat{\alpha}_2$	0.99	0.99	0.99	0.99	-0.69	0.99
$\hat{\beta}_1$	0.31	0.34	0.53	0.20	-0.22	0.69
$\hat{\beta}_2$	0.95	0.94	0.85	0.98	0.98	0.72

over the different projection indices selected.

Permutation tests for independence: As the above examples show, it is often of interest to test whether the two sets of variables are independent. This hypothesis can be tested using the value of $\hat{\rho}_R$. The influence function derived in Section 3 suggests that under normality the asymptotic distribution of the test statistic $\hat{\rho}_R$ will be the same as the distribution of (the absolute value of) the association measure R_n. The latter distributions under the null hypothesis are, certainly in case of R_S and R_K, well known. Under the hypothesis of multivariate normality, the Maximum Likelihood test is Wilks' lambda (Wilks 1932, Johnson and Wichern 1998, page 322). However, when deviating from normality this test looses its power and several more robust tests for multivariate independence have been proposed (e.g.

Gieser and Randles 1997, Taskinen, Kankainen and Oja 2003). These tests have been shown to have a larger power at heavy tailed elliptical distribution. Since we prefer not to rely on distributional assumptions, the null hypothesis will be tested by a (conditional) permutation test for independence (e.g. Good 2000).

For a data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ we generate $nperm = 1000$ data sets by randomly permuting the x_i-observations, while keeping the y_i fixed. For each of these *nperm* data sets, the test statistic $\hat{\rho}_R$ is computed. Under the null, one generates in this way a sequence of replicates from the distribution of the test statistic (conditional on the x_i values.) The critical value for the test is the 95% quantile of these simulated values. Alternatively, we can compute a p-value as the percentage of replicates of the test statistic being larger than the association measure computed from the unpermuted data.

The use of robust measures of association will result in robust test procedures. Moreover, when working with Spearman or Kendall correlation (who capture not only linear but also monotone relationships), one may expect these tests to be more powerful for detecting non-linear associations between X and Y.

Permutation tests for testing independence between the (X_1, X_2) and (Y_1, Y_2, Y_3) of the Diabetes data have been carried out. The resulting p-values are presented in Table 4 in the first row. Is it observed that most test procedures are strongly rejecting the null hypothesis. The correlation median procedure would still reject the null at a level of 5%. However, the p-value is higher compared to the other measures, witnessing its low power. Since we are also interested in robustness properties, we generate an outlier in this data set by replacing the value 0.81 of object 1 of variable Y_1 by 8.1 (similar as in Taskinen et al). In fact, outliers of this kind when the comma is wrongly placed appear frequently in practice. The resulting p-values are presented in the second row of Table 4. The test based on the first canonical correlation coefficient is very sensitive to this outlier, yielding now an insignificant result. MCD and Comedian are also somehow affected, all other p-values are very stable.

Similarly, the p-values for the "HIV data" are presented in the same Table 4. The outlier is now created by changing the first observation from $x_1 = (227, 1171)^t$ and $y_1 = (469, 2879)^t$ into $x_1 = (2227, 5171)^t$ and $y_1 = (469, 879)^t$, in the same way as was done by Randles (2000). All permutation tests strongly reject the independence assumption. When the outlier is induced in the data, the result based on the first correlation coefficient changes, although the null would still be rejected at a level of 5%. But for all other procedures, including those based on ρ_S and ρ_K, the p-values are resistant to the outlier.

Table 4: Permutation tests for independence for several measures of association for the Diabetes and HIV data. Presented are the p-values, computed for the original data and the data set with one outlier added.

	Corr	S	K	M	MCD	Comed
Diabetes data	0.000	0.000	0.000	0.000	0.005	0.019
+ one outlier	0.295	0.003	0.000	0.000	0.018	0.033
HIV data	0.000	0.000	0.000	0.000	0.003	0.000
+ one outlier	0.040	0.000	0.000	0.000	0.004	0.001

6 Simulation Experiments

The sample counterparts of the association and weighting vectors all give estimators of the population quantities ρ_R, α_R, and β_R. To allow for a comparison between these different estimators, we will work at a multivariate normal model distribution $H_0 = N(0, \Sigma)$. It was shown in Section 3 that at this model the α_R and β_R are the same for every R (satisfying conditions (i)-(iv)), hence we may simply denote them by α and β. Moreover, after applying the inverse κ_R-transformation, defined in (2.4), to the association measures, also the $\kappa_R^{-1}(\rho_R)$ coincide and are denoted by ρ. Aim is now to compare the Mean Squared Error (MSE) of the estimators $\kappa_R^{-1}(\hat{\rho}_R)$, $\hat{\alpha}_R$, and $\hat{\beta}_R$ of ρ, α and β for different projection indices R.

For $m = 1, \ldots, M = 1000$ simulations, observations $x_1^m, \ldots, x_n^m \in \mathbb{R}^p$ and $y_1^m, \ldots, y_n^m \in \mathbb{R}^q$ were generated from a specified $N(0, \Sigma)$ distribution. We selected sample size $n = 100$, $p = 1$, and $q = 2$. For these small values of p and q we checked in Section 4 that the approximative algorithm for computing the estimator is extremely precise. This has the advantage that the simulation study will measure the MSE of the estimators as defined in the paper, without too much nuisance from the algorithm being used. The matrix Σ was taken to be the identity matrix, except for $\Sigma_{12} = \Sigma_{21} = \rho = 0.5$. Besides generating data at the model distribution, we also study the robustness of the different estimators by adding outliers. Herefore we randomly replace rows of the data matrix by observations coming from a $N(\mu_c, \Sigma_c)$ distribution, with Σ_c an identity matrix except for $\Sigma_{13} = \Sigma_{31} = 1$, and $\mu_c = (\lambda, 0, \lambda)^t$. The outliers have a different correlation structure, due to the choice of Σ_c, and we will consider them to be either intermediate outliers (for $\lambda = 3$) or extreme outliers (for $\lambda = 10$). While far away outliers could be detected by applying robust outlier detection rules, this is not true anymore for intermediate outliers. In this sense, robustness w.r.t. intermediate outliers is more important than w.r.t. extreme outliers. The amount of outliers considered was ranging from 0 to 15% in steps of 5%.

The mean squared error (MSE) for the estimators of ρ is computed as

$$\mathrm{MSE}_R(\rho) = \frac{1}{M}\sum_{m=1}^{M}\left(\phi(\kappa_R^{-1}(\hat{\rho}_R^m)) - \phi(\rho)\right)^2, \tag{6.1}$$

where $\phi(\rho) = \tanh^{-1}(\rho)$ is the Fisher transformation of ρ (which is often applied to render the finite sample distribution of correlation coefficients more towards normality). For the weighting vectors the MSE is defined as

$$\mathrm{MSE}_R(\beta) = \frac{1}{M}\sum_{m=1}^{M}\left\{\cos^{-1}\left(\frac{|\beta^t\hat{\beta}_R{}^m|}{||\hat{\beta}_R{}^m||\cdot||\beta||}\right)\right\}^2. \tag{6.2}$$

The measure (6.2) is the average value of the angles between the vectors $\hat{\beta}_R{}^m$ and β. The use of angles makes the MSE invariant to the choice of the normalization constraint for the weighting vectors.

At the model distribution (in absence of outliers) the simulated MSEs are in the table below:

	Corr	S	K	M	MCD	Comed
100 $\times\mathrm{MSE}_R(\rho)$	0.991	1.147	1.157	1.084	1.879	5.159
$\mathrm{MSE}_R(\beta)$	0.035	0.049	0.052	0.043	0.101	0.173

The estimates based on the classical correlation are most precise (in fact, it are the Maximum Likelihood estimators), closely followed by the Huber M, and the estimators based on the non-parametric measures. For MCD (in fact, the reweighted version of MCD) the loss in efficiency at the model is quite important and the approach based on the correlation median is very inefficient.

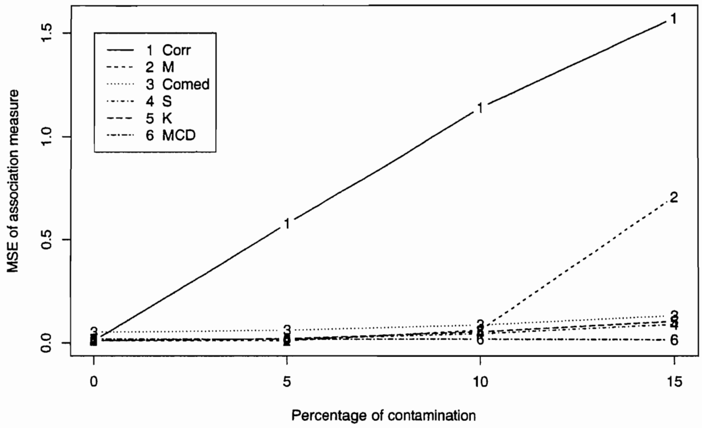
Figure 3 shows how the MSEs change in presence of extreme outliers. The level of contamination is on the horizontal axis. We clearly observe the non-robustness of the first canonical correlation coefficient, and also the M-estimator suffers for larger amounts of contamination. The MSE of the other procedures remains quite stable when increasing the level of contamination upto 15 %. In Figure 4, one sees how the MSEs increases when adding intermediate outliers. The MSE of the Corr and M based procedure increases now slower, but they are still among the less precise estimators in presence of 10 to 15% of outliers. Note that the S, K and MCD based procedures are less resistant to intermediate than to extreme outliers, certainly for the weighting vectors. At the highest considered level of contamination, the MCD performs best. But Spearman and Kendall do not loose much precision in presence of outliers, certainly not for the association measure. Given their higher efficiency at the model, one could conclude from this simulation study that $\hat{\rho}_S$ and $\hat{\rho}_K$ are very competitive, both in terms of robustness and in terms of efficiency.

In Section 5 it was shown how permutation tests can be used to test for independence between X and Y. Particularly a test using the association measure based on Spearman rank correlation could be powerful in detecting monotone, but not necessarily linear, relationships between combinations of component X and Y. In the following experiment we would like to confirm this by simulating a power curve.

We generated samples of size $n = 50$ were the X-variable is standard normal and the components of Y are generated as $Y_1 = \exp(\lambda X + \delta_1)$, and $Y_j = \exp(\delta_j)$ for $j > 1$. The δ_j variables are all independent, standard normally distributed. To keep the computation time within limits, we selected $p = 1$, $q = 2$. The null hypothesis of independence corresponds with $\lambda = 0$, while larger values of λ imply existence of non-linear relationships and should result in frequent rejection of the null hypothesis. For λ varying from 0 to 1 in steps of 0.2, empirical rejection frequencies have been computed on the basis of 1000 simulation runs. The results are represented in the form of a powercurve in Figure 5a. It is readily seen that the test based on R_S is more powerful than the test based on Corr, the difference being larger than 10% for some values of λ. A similar powercurve for alternatives with linear dependencies was simulated. Herefore the same experiment as above was done, but without carrying out the exponential transformation. The resulting powercurve in Figure 5b shows only a minor loss in efficiency of S in comparison w.r.t. Corr for this alternative.

7 Conclusions

Many different measures of association have been introduced in the statistical literature. Besides Pearson's correlation coefficient, measuring linear association, Spearman's rank measure for monotone relations between two variables is probably the best known. In this paper a projection-pursuit based approach for quantifying the association between two multivariate variables is introduced. Its definition is intuitively appealing: it is the highest possible association R that we can find between any two indices $\alpha^t X$ and $\beta^t X$ constructed from the two sets of variables X and Y. Using different measures R, different measures of association are obtained. This paper focused mainly on two topics: (i) the robustness of the measures



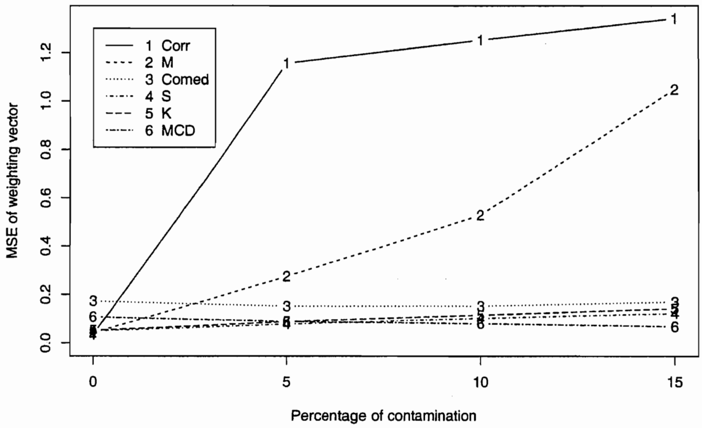
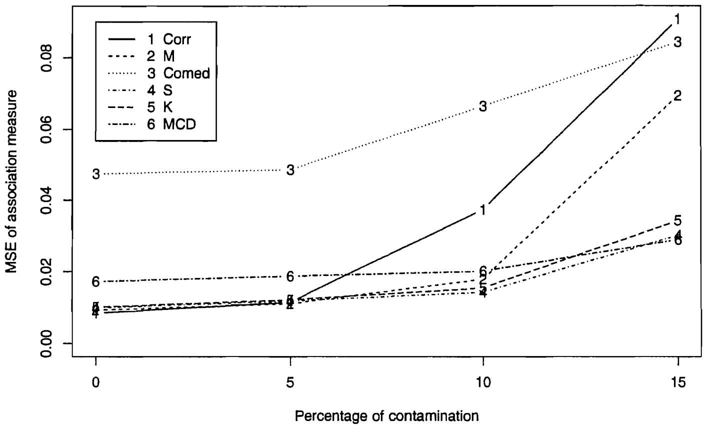


Figure 3: Simulated MSEs for the estimators of the association measure (upper figure) and of the weighting vector (lower figure) using different projection indices: Pearson correlation (Corr), Spearman's rank correlation (S), Kendall's τ (K), correlation based on Huber's bivariate M-estimator (M) and on the bivariate MCD and the correlation median (Comed). Samples were generated from a normal distribution, but contained 0 to 15% **extreme outliers**.



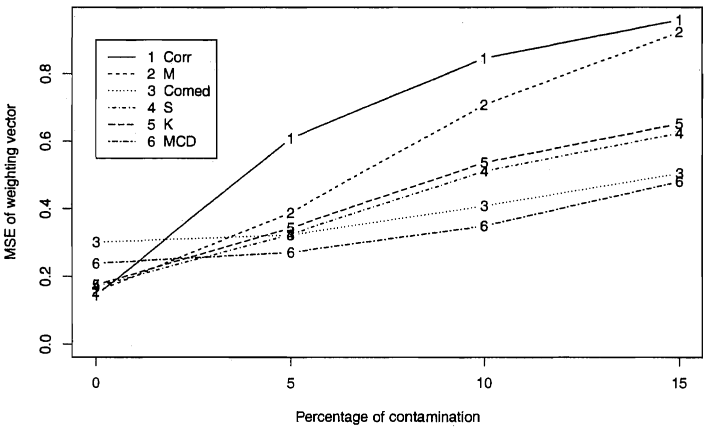
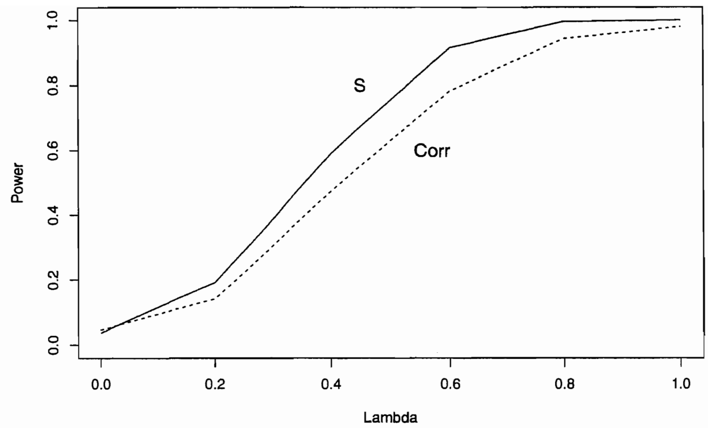


Figure 4: As Figure 3, but now samples were generated from a normal distribution, containing 0 to 15% **intermediate** *outliers.*

(a)



(b)

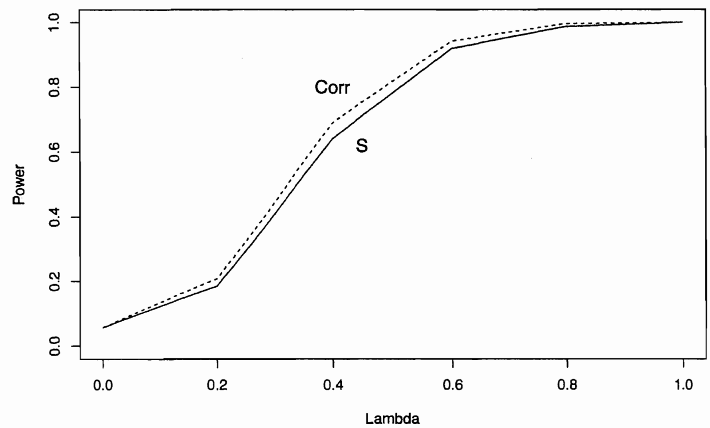


Figure 5: Powercurve of permutations tests for independence using $\hat{\rho}_R$ based on Corr and based on Spearman rank correlation (a) with respect to alternatives with non-linear dependency (upper figure), and (b) with respect to alternatives with linear dependency (bottom figure).

(ii) the use of the measures in tests for independence. For (i) influence calculations have been made, showing that the projection index R should have a bounded and smooth influence function. Moreover, the simulation study in Section 6 studied the stability of the measures under contamination. The asymptotic distribution of the different estimators has not been derived. We conjecture, however, that it is asymptotically normal. A proof could be given along the lines of Cui, He, and Ng (2003). We have a particular preference for using the Spearman rank correlation as projection index. It has good robustness properties, and for (ii) it was shown to have power for detecting non-linear relations. Moreover, it has only a small loss in efficiency and a small loss of power for testing independency at multivariate normal distributions in comparison with the approach based on the non robust Pearson correlation. Finally, using the proposed algorithm, $\hat{\rho}_S$ is fast to compute. For the examples in the paper, carrying out a permutation test with 1000 randomizations based on $\hat{\rho}_S$ took less than a second using a modern laptop computer.

Although ρ_{Corr} is the first canonical correlation coefficient, we prefer to see ρ_R as an association measure and not as an attempt to robustify canonical correlation analysis (CCA). In robust CCA (e.g. Karnel 1991 using M-estimators, Croux and Dehon 2002 using the MCD) one estimates robustly the covariance matrix of the data, and takes the eigenvalues of the robust estimate of (2.2) as robust canonical correlations. But the first canonical correlation obtained in this way is not interpretable anymore as a correlation measure between linear combinations of the components of X and Y. Only at elliptically symmetric distributions it is interpretable in this sense. Note that a small simulation study comparing CCA based on robust covariances and using a projection pursuit approach has been carried out by Dehon, Croux and Filzmoser (2000) and independently by Oliveira and Branco (2000). Furthermore, we do not aim in this paper to introduce counterparts of higher order canonical correlations. It is not obvious how a natural and interpretable definition for higher order measures of association can be obtained.

We believe that a projection-pursuit based measure of association based on, for example, Spearman's rank correlation will be of use in several applications. Besides being a descriptive measure, it has already been shown how it can be used for testing independence. Another potential application is the case where $p = 1$ and X is only measured on an ordinal scale. Then the rank of $\beta_{R_S}^t Y$ is a more natural predictor for X than $\beta_{R_{\text{Corr}}}^t Y$, the latter one being optimal in the least squares sense.

Appendix

Proof of Theorem 1. Let $(X_\varepsilon, Y_\varepsilon) \sim H_\varepsilon$, where $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{(x,y)}$. We need to solve a maximization under constraints, with associated Lagrangian function

$$\mathcal{L}(\alpha,\beta) = R(\alpha^t X_\varepsilon, \beta^t Y_\varepsilon) - \lambda_1(\varepsilon)\left[\alpha^t\alpha - 1\right] - \lambda_2(\varepsilon)\left[\beta^t\beta - 1\right]$$

for $\alpha \in I\!R^p$, $\beta \in I\!R^q$.

Let $\psi_\alpha(\varepsilon) = \frac{\partial}{\partial\alpha} R(\alpha^t X_\varepsilon, \beta^t Y_\varepsilon)_{|(\alpha_\varepsilon,\beta_\varepsilon)}$ and $\psi_\beta(\varepsilon) = \frac{\partial}{\partial\beta} R(\alpha^t X_\varepsilon, \beta^t Y_\varepsilon)_{|(\alpha_\varepsilon,\beta_\varepsilon)}$. The first order equations for the above Lagrangian function are then given by

$$\left\{\begin{array}{l} \psi_\alpha(\varepsilon) = 2\lambda_1(\varepsilon)\alpha \\ \psi_\beta(\varepsilon) = 2\lambda_2(\varepsilon)\beta. \end{array}\right. \tag{7.1}$$

We have that $\alpha_\varepsilon = \alpha_R(H_\varepsilon)$ and $\beta_\varepsilon = \beta_R(H_\varepsilon)$ are solutions of the above system. Since $R(\alpha^t X, \beta^t Y) = \kappa_R\left(r(\alpha,\beta)\right)$, see (3.3), we have that

$$\psi_\alpha(0) = \kappa_R'\left(r(\alpha,\beta)\right)\left\{\frac{\Sigma_{xy}\beta}{s(\alpha)s(\beta)} - \frac{r(\alpha,\beta)\Sigma_{xx}\alpha}{s^2(\alpha)}\right\} \tag{7.2}$$

$$\psi_\beta(0) = \kappa_R'\left(r(\alpha,\beta)\right)\left\{\frac{\Sigma_{yx}\alpha}{s(\alpha)s(\beta)} - \frac{r(\alpha,\beta)\Sigma_{yy}\beta}{s^2(\alpha)}\right\} \tag{7.3}$$

where we used the notations $s(\alpha) = \sqrt{\alpha^t\Sigma_{xx}\alpha}$ and $s(\beta) = \sqrt{\beta^t\Sigma_{yy}\beta}$.
Now let $\alpha_0 = \alpha_R(H_0)$ and $\beta_0 = \beta_R(H_0)$ be the weighting vectors at the model distribution. Then we know that α_0 and β_0 have norm 1 and satisfy (7.1) for $\varepsilon = 0$. Using (7.2) and premultiply LHS and RHS of (7.2) by α_0^t yields

$$\alpha_0^t\psi_\alpha(0) = \kappa_R'\left(r(\alpha_0,\beta_0)\right)\left\{r(\alpha_0,\beta_0) - r(\alpha_0,\beta_0\right\} = 2\lambda_1(0).$$

Hence $\lambda_1(0) = 0$ and similarly $\lambda_2(0) = 0$. Derivating (7.1) w.r.t. ε and evaluating at 0 yields therefore

$$\left\{\begin{array}{l} \frac{\partial}{\partial\varepsilon}\psi_\alpha(\varepsilon)_{|\varepsilon=0} = 2\frac{\partial}{\partial\varepsilon}\lambda_1(\varepsilon)_{|\varepsilon=0}\,\alpha_0 \\ \frac{\partial}{\partial\varepsilon}\psi_\beta(\varepsilon)_{|\varepsilon=0} = 2\frac{\partial}{\partial\varepsilon}\lambda_2(\varepsilon)_{|\varepsilon=0}\,\beta_0. \end{array}\right. \tag{7.4}$$

Premultiplying LHS and RHS of the first equation of (7.4) by α_0^t gives

$$2\frac{\partial}{\partial\varepsilon}\lambda_1(\varepsilon)_{|\varepsilon=0} = \alpha_0^t\frac{\partial}{\partial\varepsilon}\psi_\alpha(\varepsilon)_{|\varepsilon=0} \tag{7.5}$$

In the same way

$$2\frac{\partial}{\partial\varepsilon}\lambda_2(\varepsilon)_{|\varepsilon=0} = \beta_0^t\frac{\partial}{\partial\varepsilon}\psi_\beta(\varepsilon)_{|\varepsilon=0} \tag{7.6}$$

such that (7.4) implies

$$\begin{cases} (I-\alpha_0\alpha_0^t)\frac{\partial}{\partial\varepsilon}\psi_\alpha(\varepsilon)|_{\varepsilon=0}=0 \\ (I-\beta_0\beta_0^t)\frac{\partial}{\partial\varepsilon}\psi_\beta(\varepsilon)|_{\varepsilon=0}=0. \end{cases} \tag{7.7}$$

Application of the chain rule allows to compute the derivatives

$$\begin{aligned} \frac{\partial}{\partial\varepsilon}\psi_\alpha(\varepsilon)|_{\varepsilon=0} &= \frac{\partial}{\partial\alpha}\frac{\partial}{\partial\alpha^t}R(\alpha^tX,\beta^tY)|_{(\alpha_0,\beta_0)}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} \\ &+\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta^t}R(\alpha^tX,\beta^tY)|_{(\alpha_0,\beta_0)}\frac{\partial}{\partial\varepsilon}\beta_\varepsilon|_{\varepsilon=0} \\ &+\frac{\partial}{\partial\varepsilon}\left[\frac{\partial}{\partial\alpha}R(\alpha^tX_\varepsilon,\beta^tY_\varepsilon)|_{(\alpha_0,\beta_0)}\right]|_{\varepsilon=0}. \end{aligned} \tag{7.8}$$

Call these three terms T_1, T_2 and T_3. For the last term we change the order of differentiation and obtain

$$T_3=\frac{\partial}{\partial\alpha}\left[\frac{\partial}{\partial\varepsilon}R(\alpha^tX_\varepsilon,\beta^tY_\varepsilon)|_{\varepsilon=0}\right]|_{(\alpha_1,\beta_1)}=\frac{\partial}{\partial\alpha}IF((\alpha^tx,\beta^ty),R,\Phi_{\alpha,\beta}) \tag{7.9}$$

where $\Phi_{\alpha,\beta}$ is the bivariate $N\left(0,\begin{pmatrix}\alpha^t\Sigma_{xx}\alpha & \alpha^t\Sigma_{xy}\beta \\ \beta^t\Sigma_{yx}\alpha & \beta^t\Sigma_{yy}\beta\end{pmatrix}\right)$ distribution. Due to equivariance of R, (7.9) can be written as

$$T_3=\frac{\partial}{\partial\alpha}(\alpha^t\Sigma_{xx}\alpha)IF\left(\left(\frac{\alpha^tx}{\sqrt{\alpha^t\Sigma_{xx}\alpha}},\frac{\beta^ty}{\sqrt{\beta^t\Sigma_{yy}\beta}}\right),R,\Phi_\rho\right)|_{(\alpha_0,\beta_0)},$$

which can be checked to equal (using $\alpha_0=\alpha_1s(\alpha_0)$)

$$T_3=IF_1\left((u_1,v_1),R,\Phi_\rho\right)\frac{P_{\alpha_1}^{\perp}x}{s(\alpha_0)},$$

with $u_1=\alpha_1^tx$ and $v_1=\beta_1^ty$. Here we used the notations $P_{\alpha_1}^{\perp}=(I-\Sigma_{xx}\alpha_1\alpha_1^t)$ and

$$IF_1\left((u_1,v_1),R,\Phi_\rho\right)=\frac{\partial}{\partial u_1}IF\left((u_1,v_1),R,\Phi_\rho\right).$$

Now for the first term of (7.8) we have

$$T_1=\frac{\partial}{\partial\alpha^t}\psi_\alpha(0)|_{(\alpha_0,\beta_0)}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0}. \tag{7.10}$$

Using (7.2), condition (iv) and the fact that $\psi_\alpha(0)=2\lambda_1(0)\alpha_0=0$ (cf. (7.1) and $\lambda_1(0)=0$), we have

$$\begin{aligned} \frac{\partial}{\partial\alpha^t}\psi_\alpha(0) &= \kappa_R'\left(r(\alpha_0,\beta_0)\right)\frac{\partial}{\partial\alpha^t}\left\{\frac{1}{s(\alpha)}\left(\Sigma_{xy}\frac{\beta}{s(\beta)}-\frac{r(\alpha,\beta)\Sigma_{xx}\alpha}{s(\alpha)}\right)\right\}|_{(\alpha_0,\beta_0)} \\ &= -\kappa_R'(\rho_1)\frac{\partial}{\partial\alpha^t}\left\{\frac{\alpha^t\Sigma_{xy}\beta}{s(\beta)s(\alpha)^2}\Sigma_{xx}\alpha\right\}|_{(\alpha_0,\beta_0)}. \end{aligned}$$

Careful matrix differentiation, combined with using (3.5) at (α_0, β_0) yields then,

$$T_2 = -\frac{\rho_1 \kappa_R'(\rho_1)}{s^2(\alpha_0)} P_{\alpha_1}^{\perp} \Sigma_{xx} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0}. \tag{7.11}$$

Following similar calculations, we end up for the second term of (7.8) with

$$T_2 = \frac{\kappa_R'(\rho_1)}{s(\alpha_0)s(\beta_0)} P_{\alpha_1}^{\perp} \Sigma_{xy} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0}. \tag{7.12}$$

The first equation of (7.7), together with $(I - \alpha_0 \alpha_0^t) P_{\alpha_1}^{\perp} = P_{\alpha_1}^{\perp}$ since α_0 and α_1 are parallel, yields

$$\begin{aligned} 0 &= (I - \alpha_0\alpha_0^t)(T_1 + T_2 + T_3) = \\ & -\frac{\rho_1 \kappa_R'(\rho_1)}{s^2(\alpha_0)} P_{\alpha_1}^{\perp} \Sigma_{xx} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} + \frac{\kappa_R'(\rho_1)}{s(\alpha_0)s(\beta_0)} P_{\alpha_1}^{\perp} \Sigma_{xy} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} + IF_1((u_1, v_1), R, \Phi_\rho) P_{\alpha_1}^{\perp} x, \end{aligned} \tag{7.13}$$

from which it follows that

$$\kappa_R'(\rho_1) P_{\alpha_1}^{\perp} \left(\frac{\Sigma_{xy}}{s(\beta_0)} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} - \frac{\rho_1 \Sigma_{xx}}{s(\alpha_0)} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} \right) = -IF_1((u_1, v_1), R, \Phi_\rho) P_{\alpha_1}^{\perp} x. \tag{7.14}$$

By symmetry, one also has

$$\kappa_R'(\rho_1) P_{\beta_1}^{\perp} \left(\frac{\Sigma_{yx}}{s(\alpha_0)} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} - \frac{\rho_1 \Sigma_{yy}}{s(\beta_0)} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} \right) = -IF_2((u_1, v_1), R, \Phi_\rho) P_{\beta_1}^{\perp} y. \tag{7.15}$$

Let $\alpha_2, \alpha_3, \ldots, \alpha_p$ and $\beta_2, \ldots, \beta_q$ be the population higher order canonical vectors associated with H. Suppose $q > p$ w.l.o.g. The higher order canonical correlations are $\lambda_2, \ldots, \lambda_p$ (where the last $q - p$ are set equal to zero). Then $\beta_1, \ldots, \beta_q$ form an orthonormal basis for the inner product imposed by Σ_{yy}. So we have $I = \sum_{j=1}^q \beta_j \beta_j^t \Sigma_{yy}$ such that (7.15) can be rewritten as

$$\kappa_R'(\rho_1) \sum_{j=2}^q \Sigma_{yy} \beta_j \beta_j^t \Sigma_{yy} \left(\frac{\Sigma_{yy}^{-1} \Sigma_{yx}}{s(\alpha_0)} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} - \frac{\rho_1}{s(\beta_0)} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} \right) = -IF_2((u_1, v_1), R, \Phi_\rho) \sum_{j=2}^q \Sigma_{yy} \beta_j \beta_j^t y.$$

Now premultiplying the above equation by β_k^t, for a fixed $1 < k \leq p$ $(\leq q)$ and using $\rho_k \Sigma_{yy} \beta_k = \Sigma_{yx} \alpha_k$ gives

$$\kappa_R'(\rho_1) \alpha_k^t \Sigma_{xy} \left(\frac{\Sigma_{yy}^{-1} \Sigma_{yx}}{s(\alpha_0)} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} - \frac{\rho_1}{s(\beta_0)} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} \right) = -IF_2((u_1, v_1), R, \Phi_\rho) \beta_k^t y,$$

or, since α_k is an eigenvector of $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}$ with eigenvalue ρ_k^2, and with $v_k = \beta_k^t y$,

$$\kappa_R'(\rho_1) \alpha_k^t \frac{\Sigma_{xy}}{s(\beta_0)} \frac{\partial}{\partial \varepsilon} \beta_\varepsilon |_{\varepsilon=0} = \kappa_R'(\rho_1) \frac{\rho_k^2}{\rho_1} \alpha_k^t \frac{\Sigma_{xx}}{s(\alpha_0)} \frac{\partial}{\partial \varepsilon} \alpha_\varepsilon |_{\varepsilon=0} + \frac{\rho_k}{\rho_1} IF_2((u_1, v_1), R, \Phi_\rho) v_k. \tag{7.16}$$

Premultiplying (7.14) now by α_k^t yields then, since $\alpha_k^t P_{\alpha_1}^{\perp} = \alpha_k^t$, and by plugging in (7.16):

$$\kappa_R'(\rho_1)\left(\frac{\rho_k^2}{\rho_1}\alpha_k^t\frac{\Sigma_{xx}}{s(\alpha_0)}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} - \rho_1\alpha_k^t\frac{\Sigma_{xx}}{s(\alpha_0)}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0}\right)$$
$$= -\frac{\rho_k}{\rho_1}IF_2((u_1,v_1),R,\Phi_\rho)v_k - IF_1((u_1,v_1),R,\Phi_\rho)u_k$$

with $u_k = \alpha_k^t x$. We conclude that

$$\alpha_k^t\Sigma_{xx}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} = \frac{1}{\rho_1^2-\rho_k^2}[\rho_1 u_k IF_1((u_1,v_1),R,\Phi_\rho) + \rho_k v_k IF_2((u_1,v_1),R,\Phi_\rho)]\frac{s(\alpha_0)}{\kappa_R'(\rho_1)} \quad (7.17)$$

for every $1 < k \leq p$. Now derivation of the side condition $\alpha_\varepsilon^t\alpha_\varepsilon = 1$ yields

$$\alpha_0^t\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} = 0, \quad (7.18)$$

hence

$$\begin{aligned} IF((x,y),\alpha,H) &= \frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} = (I-\alpha_0\alpha_0^t)\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} = (I-\alpha_0\alpha_0^t)\sum_{k=1}^p\left(\alpha_k^t\Sigma_{xx}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0}\right)\alpha_k \\ &= \sum_{k=2}^p\left(\alpha_k^t\Sigma_{xx}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0}\right)(I-\alpha_0\alpha_0^t)\alpha_k. \end{aligned}$$

The above equation, together with (7.17) and $\alpha_0 = \alpha_1/\|\alpha_1\|$ imply (3.8). The expression for $IF((x,y),b,H)$ follows in an analogous way, starting from equations (7.14) and (7.15).

Finally, since $\rho_R(H_\varepsilon) = R(\alpha_\varepsilon^t X_\varepsilon, \beta_\varepsilon^t Y_\varepsilon)$, using (7.18), we have

$$\begin{aligned} \frac{\partial}{\partial\varepsilon}\rho_R(H_\varepsilon)|_{\varepsilon=0} &= \frac{\partial}{\partial\alpha}R(\alpha^tX,\beta^tY)|_{(\alpha_0,\beta_0)}\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} + \frac{\partial}{\partial\beta}R(\alpha^tX,\beta^tY)|_{(\alpha_0,\beta_0)}\frac{\partial}{\partial\varepsilon}\beta_\varepsilon|_{\varepsilon=0} \\ &\quad + \frac{\partial}{\partial\varepsilon}R(\alpha_0^tX_\varepsilon,\beta_0^tY_\varepsilon)|_{\varepsilon=0} \\ &= \psi_\alpha(0)\frac{\partial}{\partial\varepsilon}\alpha_\varepsilon|_{\varepsilon=0} + \psi_\beta(0)\frac{\partial}{\partial\varepsilon}\beta_\varepsilon|_{\varepsilon=0} + IF((\alpha_0^tx,\beta_0^ty),R,\Phi_{(\alpha_0,\beta_0)}) \\ &= IF((u_1,v_1),R,\Phi_\rho) \end{aligned}$$

where we used $\psi_\alpha(0) = \psi_\beta(0) = 0$. □

8 References

Andrews, D.F., and Herzberg, A.M. (1985). *Data.* Springer-Verlag, New York.

Boente, G., Pires, A.M., and Rodrigues, I.M. (2002). Influence functions and outlier detection under the common principal components model: a robust approach. *Biometrika*, 89, 861–875.

Butler, R.W., Davies, P.L., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21, 1385–1400.

Croux, C., and Dehon, C. (2002). Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. *La Revue de Statistique Appliquee*, 2, 5-26

Croux, C., and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71, 161–190.

Croux, C., and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87, 603–618.

Croux, C., and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In: A. Prat (ed.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 211–216.

Cui, H., He, X., and Ng, K.W. (2003). Asymptotic distributions of principal components based on robust dispersions. *Biometrika*. To appear.

Dehon, C., Filzmoser, P., and Croux, C. (2000). Robust methods for canonical correlation analysis. In: J.G. Betlehem and P.G.M. van der Heijden (eds.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 321–326.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 531–546.

Falk, M. (1998). A note on the correlation median for elliptical distributions. *Journal of Multivariate Analysis*, 67, 306–317.

Gieser, P.W., and Randles, R.H. (1997). A Nonparametric Test of Independence Between Two Vectors. *Journal of the American Statistical Association*, 92, 561-567.

Good, P. (2000). *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses*. 2nd ed., Springer-Verlag, New York.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based in Influence Functions.* Wiley & Sons, New York.

Huber, P.J. (1985). Projection pursuit. *Ann. Statist.*, 13, 435–525.

Johnson, R.A., and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis.* 4th ed., Prentice-Hall, London.

Karnel, G. (1991). Robust canonical correlation and correspondence analysis. *The Frontiers of Statistical Scientific and Industrial Applications*, (Volume II of the proceedings of ICOSCO-I, The First International Conference on Statistical Computing), 335–354.

Li, G., and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80, 759–766.

Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51–67.

Maronna, R.A., and Yohai, V.J. (1998). Robust estimation of multivariate location and scatter. In: S. Kotz, C. Read, and D. Banks (eds.), *Encyclopedia of Statistical Sciences.* Wiley & Sons, New York, pp. 589-596.

Oliveira, M.R., and Branco, J.A. (2000). Projection pursuit approach to robust canonical correlation analysis. In: J.G. Betlehem and P.G.M. van der Heijden (eds.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 415–420.

Randles, R.H. (2000). A simple, affine-invariant, multivariate, distribution-free sign test. *Journal of the American Statistical Association*, 95, 1263–1268.

Rencher, A.C. (1998). *Multivariate Statistical Inference and Applications.* Wiley & Sons, New York.

Romanazzi, M. (1992). Influence in canonical correlation analysis. *Psychometrika*, 57, 237–259.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In: W. Grossmann et al. (eds.), *Mathematical Statistics and Applications, Vol. B*, Reidel, Dordrecht, pp. 283–297.

Rousseeuw, P.J., and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Shevlyakov, G.L., and Vilchevski, N.O. (2002). *Robustness in Data Analysis: Criteria and Methods.* VSP, Utrecht.

Taskinen, S., Kankainene, A., and Oja, H. (2003). Sign test of Independence Between Two Random Vectors. *Statistics and Probability Letters,* 62, 9-21.

S-Plus 2000 Guide to Statistics, Volume 1 (1999). Data Analysis Products Division, Mathsoft, Seattle, WA.

Wilks, S.S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471–494.